SMDM PROJECT

BUSINESS REPORT

Submitted By:

Prachi Gupta

Problem #1

A wholesale distributor operating in different regions of Portugal has information on annual spending of several items in their stores across different regions and channels. The data (Wholesale Customer.csv) consists of 440 large retailers' annual spending on 6 different varieties of products in 3 different regions (Lisbon, Oporto, Other) and across different sales channel (Hotel, Retail).

1.1 a) Use methods of descriptive statistics to summarize data.

To get insights into data, I have used '.**read_csv()**' function from pandas library, and to fetch first or last five observations, '.**head()**' or '.**tail()**' functions can be used. I have used '.**head()**' here.

	Buyer/Spender	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
0	1	Retail	Other	12669	9656	7561	214	2674	1338
1	2	Retail	Other	7057	9810	9568	1762	3293	1776
2	3	Retail	Other	6353	8808	7684	2405	3516	7844
3	4	Hotel	Other	13265	1196	4221	6404	507	1788
4	5	Retail	Other	22615	5410	7198	3915	1777	5185

I have used '**describe()**' function in pandas to get descriptive statistics summary. It returns the count, mean, standard deviation, minimum, maximum values and the 3 quantiles of the dataframe provided.

	Delicatessen	Detergents_Paper	Fresh	Frozen	Grocery	Milk
count	440.000000	440.000000	440.000000	440.000000	440.000000	440.000000
mean	1524.870455	2881.493182	12000.297727	3071.931818	7951.277273	5796.265909
std	2820.105937	4767.854448	12647.328865	4854.673333	9503.162829	7380.377175
min	3.000000	3.000000	3.000000	25.000000	3.000000	55.000000
25%	408.250000	256.750000	3127.750000	742.250000	2153.000000	1533.000000
50%	965.500000	816.500000	8504.000000	1526.000000	4755.500000	3627.000000
75%	1820.250000	3922.000000	16933.750000	3554.250000	10655.750000	7190.250000
max	47943.000000	40827.000000	112151.000000	60869.000000	92780.000000	73498.000000

Here, as you can see, mean value is greater than median value for all the columns showing Items, which is shown as 50% (50th percentile) in the index column.

On the basis of above shown calculation using '.**Groupby()**' method, we can conclude that:

```
df.groupby('Channel').agg({'Total':'sum'})
```

	Total
Channel	
Hotel	7999569
Retail	6619931

```
df.groupby('Region').agg({'Total':'sum'})
```

	Total
Region	
Lisbon	2386813
Oporto	1555088
Other	10677599

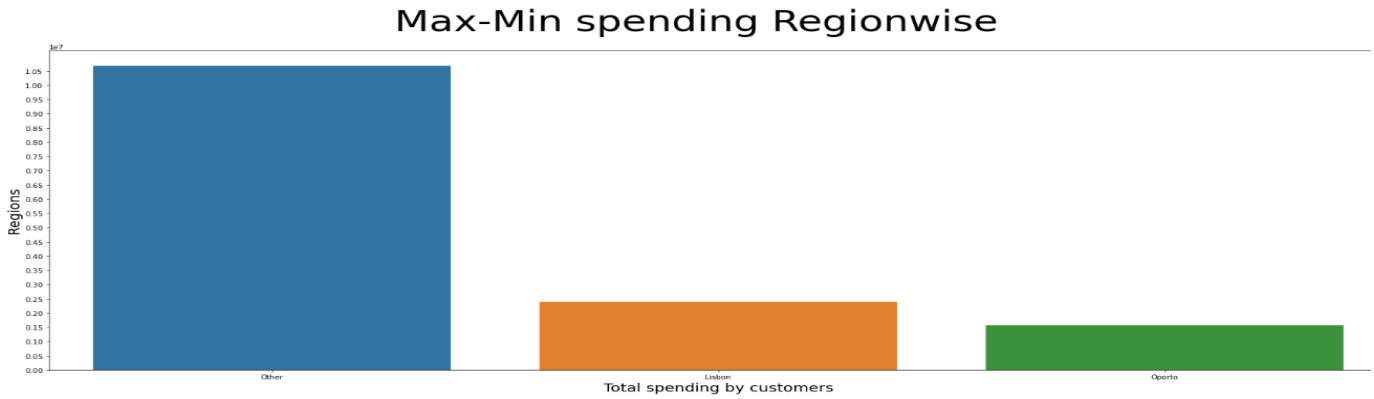
1.1 b) Which Region and which Channel spend the most:

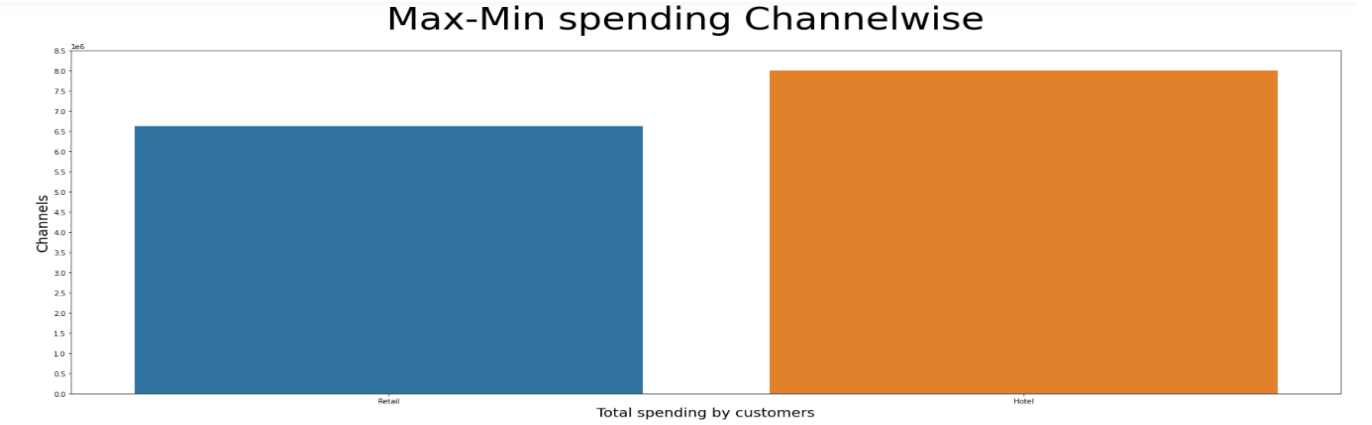
Region: Other, **Channel**: Hotel

1.1 c) Which Region and which Channel spend the least:

Region: Oporto, **Channel**: Retail

Following are the Barplots showing the max-min spending of the customers across Regions & across Channels:





1.2 There are 6 different varieties of items that are considered. Describe and comment/explain all the varieties across Region and Channel? Provide a detailed justification for your answer.

Performed the descriptive analysis of the 6 varieties, individually

Across 3 '**Region'**: Lisbon, Oporto, Other

Across 2 '**Channel'** : Hotel, Retail using describe() and Histogram plot.

Following are the descriptive summaries of the 6 varieties across 'Channel' and 'Region' :

```
df1.groupby('Channel').describe().transpose()
```

Channel		Hotel	Retail
Fresh	count	298.000000	142.000000
	mean	13475.560403	8904.323944
	std	13831.687502	8987.714750
	min	3.000000	18.000000
	25%	4070.250000	2347.750000
	50%	9581.500000	5993.500000
	75%	18274.750000	12229.750000
	max	112151.000000	44466.000000
Milk	count	298.000000	142.000000
	mean	3451.724832	10716.500000
	std	4352.165571	9679.631351
	min	55.000000	928.000000
	25%	1164.500000	5938.000000
	50%	2157.000000	7812.000000
	75%	4029.500000	12162.750000
	max	43950.000000	73498.000000
Grocery	count	298.000000	142.000000
	mean	3962.137584	16322.852113
	std	3545.513391	12267.318094
	min	3.000000	2743.000000
	25%	1703.750000	9245.250000
	50%	2684.000000	12390.000000
	75%	5076.750000	20183.500000
	max	21042.000000	92780.000000
Frozen	count	298.000000	142.000000
	mean	3748.251678	1652.612676
	std	5643.912500	1812.803662
	min	25.000000	33.000000
	25%	830.000000	534.250000
	50%	2057.500000	1081.000000
	75%	4558.750000	2146.750000
	max	60869.000000	11559.000000
Detergents_Paper	count	298.000000	142.000000
	mean	790.560403	7269.507042
	std	1104.093673	6291.089697
	min	3.000000	332.000000
	25%	183.250000	3683.500000
	50%	385.500000	5614.500000
	75%	899.500000	8662.500000
	max	6907.000000	40827.000000
Delicatessen	count	298.000000	142.000000
	mean	1415.956376	1753.436620
	std	3147.426922	1953.797047
	min	3.000000	3.000000
	25%	379.000000	566.750000
	50%	821.000000	1350.000000
	75%	1548.000000	2156.000000
	max	47943.000000	16523.000000

```
df2.groupby('Region').describe().transpose()
```

Region		Lisbon	Oporto	Other
Fresh	count	77.000000	47.000000	316.000000
	mean	11101.727273	9887.680851	12533.471519
	std	11557.438575	8387.899211	13389.213115
	min	18.000000	3.000000	3.000000
	25%	2806.000000	2751.500000	3350.750000
	50%	7363.000000	8090.000000	8752.500000
	75%	15218.000000	14925.500000	17406.500000
	max	56083.000000	32717.000000	112151.000000
Milk	count	77.000000	47.000000	316.000000
	mean	5486.415584	5088.170213	5977.085443
	std	5704.856079	5826.343145	7935.463443
	min	258.000000	333.000000	55.000000
	25%	1372.000000	1430.500000	1634.000000
	50%	3748.000000	2374.000000	3684.500000
	75%	7503.000000	5772.500000	7198.750000
	max	28326.000000	25071.000000	73498.000000
Grocery	count	77.000000	47.000000	316.000000
	mean	7403.077922	9218.595745	7896.363924
	std	8496.287728	10842.745314	9537.287778
	min	489.000000	1330.000000	3.000000
	25%	2046.000000	2792.500000	2141.500000
	50%	3838.000000	6114.000000	4732.000000
	75%	9490.000000	11758.500000	10559.750000
	max	39694.000000	67298.000000	92780.000000
Frozen	count	77.000000	47.000000	316.000000
	mean	3000.337662	4045.361702	2944.594937
	std	3092.143894	9151.784954	4260.126243
	min	61.000000	131.000000	25.000000
	25%	950.000000	811.500000	664.750000
	50%	1801.000000	1455.000000	1498.000000
	75%	4324.000000	3272.000000	3354.750000
	max	18711.000000	60869.000000	36534.000000
Detergents_Paper	count	77.000000	47.000000	316.000000
	mean	2651.116883	3687.468085	2817.753165
	std	4208.462708	6514.717668	4593.051613
	min	5.000000	15.000000	3.000000
	25%	284.000000	282.500000	251.250000
	50%	737.000000	811.000000	856.000000
	75%	3593.000000	4324.500000	3875.750000
	max	19410.000000	38102.000000	40827.000000
Delicatessen	count	77.000000	47.000000	316.000000
	mean	1354.896104	1159.702128	1620.601266
	std	1345.423340	1050.739841	3232.581660
	min	7.000000	51.000000	3.000000
	25%	548.000000	540.500000	402.000000
	50%	806.000000	898.000000	994.000000
	75%	1775.000000	1538.500000	1832.750000
	max	6854.000000	5609.000000	47943.000000

Following are the observations:

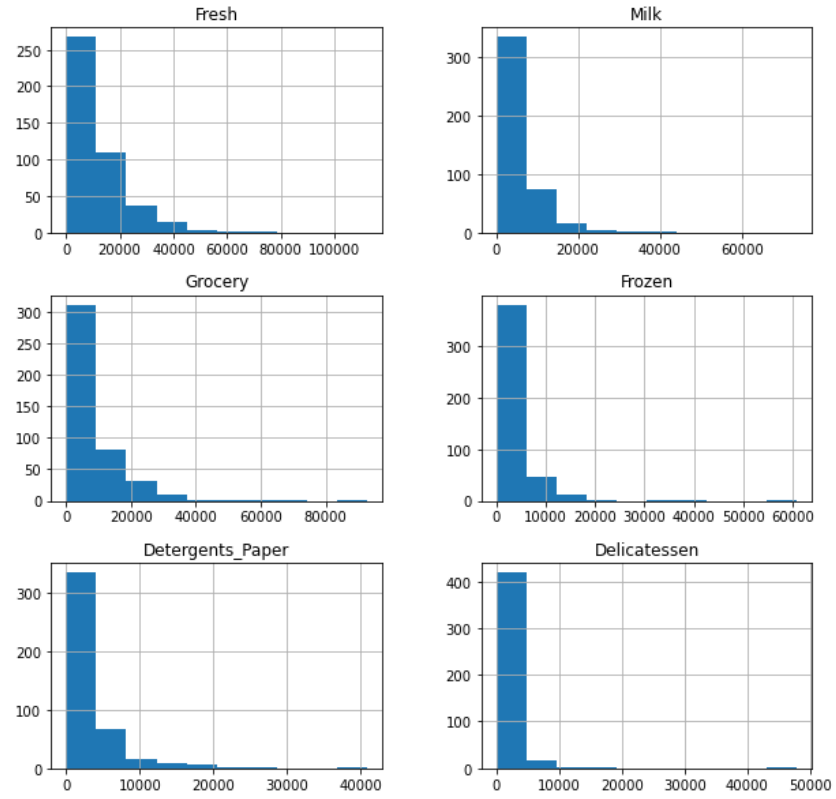
1. There are large differences between 75th %tile values and the max values for all 6 varieties across '**Region'** as well as '**Channel'** descriptive summaries, which means that there are extreme/unusual spending patterns during some times of the year.

2. The varieties: '**Fresh' , 'Milk'** and **'Grocery'** have greater Range (differences between their minimum and maximum values). Hence, the datapoints for these varieties are spread out largely.This implies that the customers buying such items have inconsistent spending patterns annually.

3. Among the 6 varieties of items across '**Channel'** and '**Region'** , the variety: '**Fresh'** has the highest mean as well as median values amidst all other varieties.Hence, this variety has the highest consumption.

4. Since the Inter-Quartile Range (Difference between 75th and 25th percentiles) is relatively very less for the varieties: '**Frozen' , 'Detergents_Paper' , 'Delicatessen' ,** the data points for these 3 varieties are more bunched up around the mean, i.e., they are very less spread out.Hence, these varieties show the least inconsistent behaviour.

5. On plotting histogram, it was observed that all 6 varieties have right-skewed/right-tailed distributions i.e., None of the distributions are symmetric about their mean values.Hence, most of the spendings are on the higher side as compared to their mean spending.




```
skewness = df1.skew(skipna=True)
skewness
```

```
Fresh              2.561323
Milk               4.053755
Grocery            3.587429
Frozen             5.907986
Detergents_Paper   3.631851
Delicatessen      11.151586
```

The same can be seen here on calculating skewness across all 6 columns.Positive values indicate right-skewed distributions.The column '**Delicatessen'** is highly positively skewed, as compared to other varieties.

6. There are some pairs of varieties which show the maximum co-variance.

   For example, '**Fresh'** has highest degree of covariance with '**Frozen'** ,which Implies that customer gener ally spending on '**Fresh'** also prefers **'Frozen'.**

   Also, '**Fresh'** has lowest degree of covariance with '**Detergents_Paper'** ,which Implies that customer ge nerally spending on '**Fresh'** is very less likely to prefer '**Detergents_Paper'.**

	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
Fresh	1.599549e+08	9.381789e+06	-1.424713e+06	2.123665e+07	-6.147826e+06	8.727310e+06
Milk	9.381789e+06	5.446997e+07	5.108319e+07	4.442612e+06	2.328834e+07	8.457925e+06
Grocery	-1.424713e+06	5.108319e+07	9.031010e+07	-1.854282e+06	4.189519e+07	5.507291e+06
Frozen	2.123665e+07	4.442612e+06	-1.854282e+06	2.356785e+07	-3.044325e+06	5.352342e+06
Detergents_Paper	-6.147826e+06	2.328834e+07	4.189519e+07	-3.044325e+06	2.273244e+07	9.316807e+05
Delicatessen	8.727310e+06	8.457925e+06	5.507291e+06	5.352342e+06	9.316807e+05	7.952997e+06

1.3 On the basis of a descriptive measure of variability, which item shows the most inconsistent behaviour? Which items show the least inconsistent behaviour?

The 3 feasible descriptive measures of variability are: IQR, Standard deviation, Coefficient of Variance.

In order to find the most/least inconsistent behaviour, we need to determine the nature of dataset first.

	IQR
Fresh	13806.00
Milk	5657.25
Grocery	8502.75
Frozen	2812.00
Detergents_Paper	3665.25
Delicatessen	1412.00

The distribution is right skewed, as skewness of all columns comes out as positive.

For a skewed dataset, having larger spread, IQR(=75th percentile -25th percentile) is a better measure to determine behaviour. Higher the IQR, more is the **inconsistent behaviour for that column.**

From the IQR calculation for all 6 item datasets, it can be concluded that:

Item having **most inconsistent** behaviour **:** '**Fresh'** (IQR=13806.00)

Item having **least inconsistent** behaviour **:** '**Delicatessen**' (IQR = 1412.00)

1.4 Are there any outliers in the data? Back up your answer with a suitable plot/technique with the help of detailed comments.

As already concluded from descriptive statistics summary, There are large differences between 75th %tile values and the max values for all 6 varieties across '**Region'** as well as '**Channel'** descriptive summaries, which means that there are Outliers/Extreme values in our dataset.

Also, Determined using the formula: Max value > IQR*1.5 , Min value < IQR*1.5

```
IQR_criteria = df_3['IQR'] *1.5
IQR_criteria
```

```
Fresh               20709.000
Milk                 8485.875
Grocery             12754.125
Frozen               4218.000
Detergents_Paper     5497.875
Delicatessen         2118.000
Name: IQR, dtype: float64
```

```
Max_Values = df.iloc[:,3:9].max()
Max_Values
```

```
Fresh               112151
Milk                 73498
Grocery              92780
Frozen               60869
Detergents_Paper     40827
Delicatessen         47943
dtype: int64
```

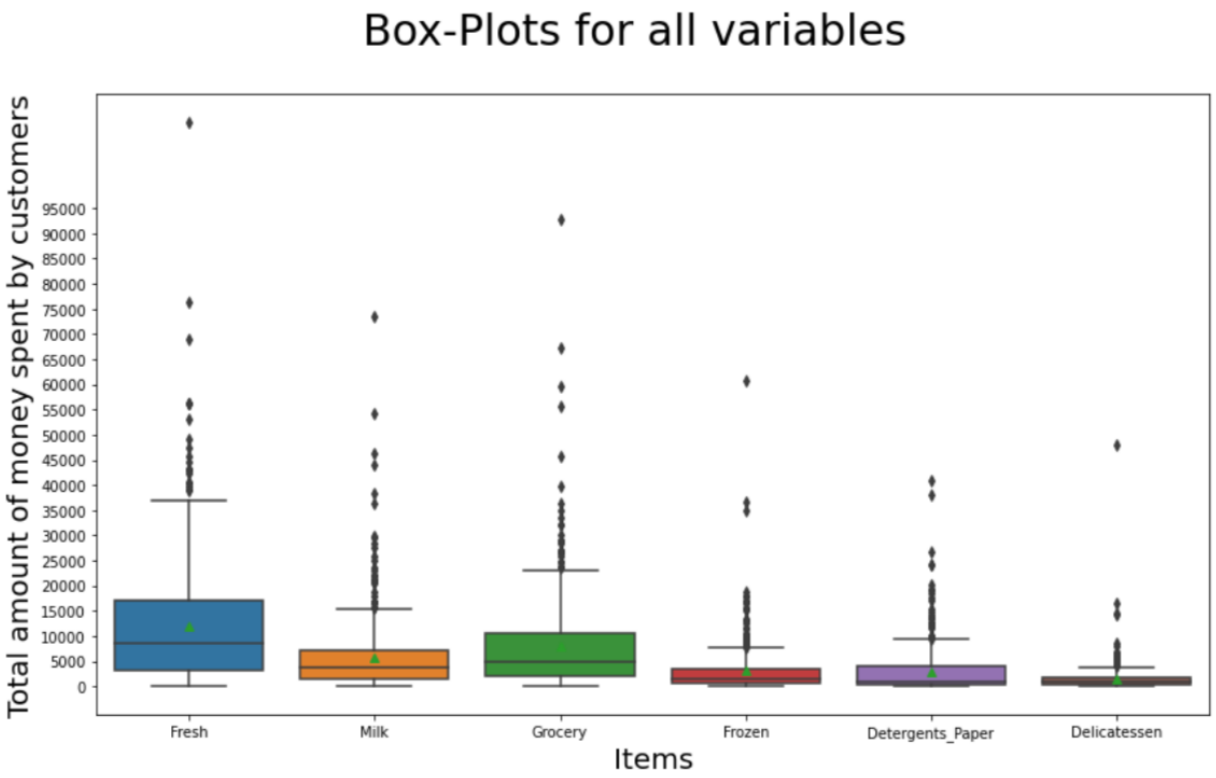
```
Min_Values = df.iloc[:,3:9].min()
Min_Values
```

```
Fresh               3
Milk               55
Grocery             3
Frozen             25
Detergents_Paper    3
Delicatessen        3
dtype: int64
```

It can be concluded from the calculations that outliers are present in our dataset.

Same is illustrated using Box-Plot below:
It shows Mean, 25%(Q1), Median(Q2), 75%(Q3) values for each column.

The upper & lower whiskers show the maximum & minimum values for that column respectively.The dotted values above the Upper Whisker represent the 'Outliers' or Extreme values.



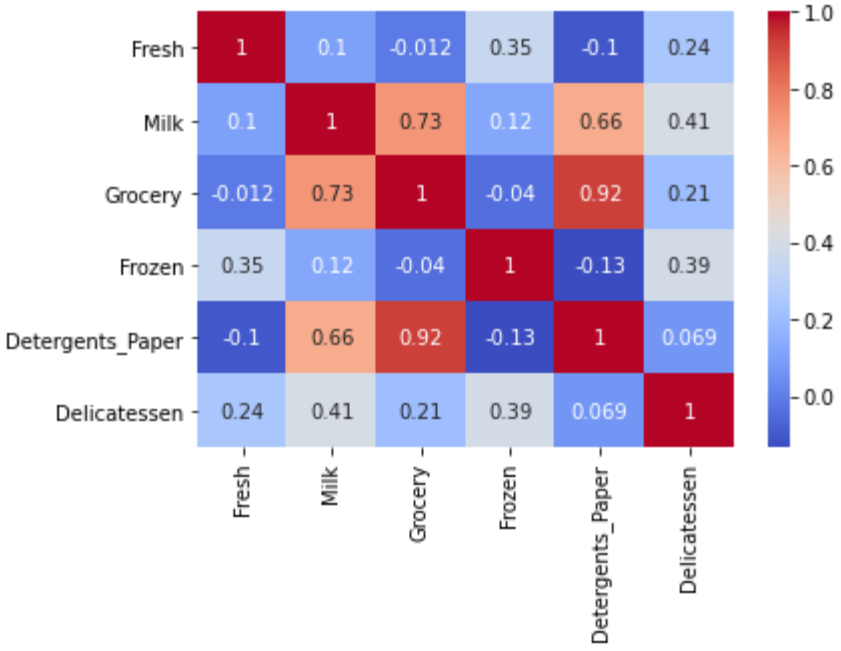
1.5 On the basis of your analysis, what are your recommendations for the business? How can your analysis help the business to solve its problem? Answer from the business perspective

Calculated the correlation among the 6 varieties first, then plotted the correlation dataframe using Heatmap.

```
df_corr = df.iloc[:,3:9].corr(method ='pearson')
df_corr
```

	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
Fresh	1.000000	0.100510	-0.011854	0.345881	-0.101953	0.244690
Milk	0.100510	1.000000	0.728335	0.123994	0.661816	0.406368
Grocery	-0.011854	0.728335	1.000000	-0.040193	0.924641	0.205497
Frozen	0.345881	0.123994	-0.040193	1.000000	-0.131525	0.390947
Detergents_Paper	-0.101953	0.661816	0.924641	-0.131525	1.000000	0.069291
Delicatessen	0.244690	0.406368	0.205497	0.390947	0.069291	1.000000

```
sns.heatmap(df_corr,annot=True , cmap='coolwarm');
```



Business Insights :

1. All varieties have extreme values, which clearly shows the unusual spending patterns across the 'Region' and 'Channel' . Also,Mean values for all the 6 varieties are greater than the median values(50th percentile) across '**Region**' as well as '**Channel'** descriptive summaries. This implies that most of the spending amount data is more than the mean values.

   It is recommended that : The annual data needs to be further drilled down or segregated on the basis of seasons, months or quarter-wise because the outliers could be due to weather changes, festivals or any other factor. For example,

   - During festivities , the demand for 'Grocery', 'Delicatessen' increases.
   - In the current COVID-19 times, demand for items under 'Detergents_Paper' like sanitizer,masks have increased drastically.

   Hence, the wholesale stores can be more prepared to stock their inventories during such times, by keeping buffer stocks to handle unusual demands.

2. Some varieties have strong/weak correlation between them, as shown in the heatmap plot above , For example,

   - 'Detergents_Paper' and 'Grocery' have a coefficient variation of 0.92
   - 'Grocery' and 'Milk' have a coefficient variation of 0.73

- ‘Grocery’ and ‘Fresh’ have a coefficient variation of -0.012
- ‘Grocery’ and ‘Frozen’ have a coefficient variation of -0.04

As concluded from the descriptive summary, the spending on items : ‘Grocery’ , ‘Milk’ and ‘Fresh’ is higher as compared to other items.

It is recommended that : The sales of items having lesser correlations with the items: ‘Grocery’ , ‘Milk’ or ‘Fresh’ can be increased by giving extra discounts or additional offers . Hence, cross selling can be promoted by providing offers on such combinations, For example,

On purchase of ‘Frozen’ or ‘Fresh’ along with ‘Grocery’ should have additional offers.

Problem #2

The Student News Service at Clear Mountain State University (CMSU) has decided to gather data about the undergraduate students that attend CMSU. CMSU creates and distributes a survey of 14 questions and receives responses from 62 undergraduates (stored in the *Survey* data set).

2.1. For this data, construct the following contingency tables (Keep Gender as row variable)

2.1.1. Gender and Major

Major Gender	Accounting	CIS	Economics/Finance	International Business	Management	Other	Retailing/Marketing	Undecided	All
Female	3	3	7	4	4	3	9	0	33
Male	4	1	4	2	6	4	5	3	29
All	7	4	11	6	10	7	14	3	62

2.1.2. Gender and Grad Intention

Grad Intention Gender	No	Undecided	Yes	All
Female	9	13	11	33
Male	3	9	17	29
All	12	22	28	62

2.1.3. Gender and Employment

Employment Gender	Full-Time	Part-Time	Unemployed	All
Female	3	24	6	33
Male	7	19	3	29
All	10	43	9	62

2.1.4. Gender and Computer

Computer	Desktop	Laptop	Tablet	All
Gender				
Female	2	29	2	33
Male	3	26	0	29
All	5	55	2	62

2.2. Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question:

2.2.1. What is the probability that a randomly selected CMSU student will be male?

Prob_RandomStudentIsMale = MaleCount/TotalCount = 29/62

2.2.2. What is the probability that a randomly selected CMSU student will be female?

Prob_RandomStudentIsFemale = FemaleCount/TotalCount = 33/62

2.3. Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question:

Major	Accounting	CIS	Economics/Finance	International Business	Management	Other	Retailing/Marketing	Undecided	All
Gender									
Female	3	3	7	4	4	3	9	0	33
Male	4	1	4	2	6	4	5	3	29
All	7	4	11	6	10	7	14	3	62

2.3.1. Find the conditional probability of different majors among the male students in CMSU.

Based on the above shown table, Following are Conditional probabilities for males:

Prob_MaleAccounting = CountOf_MalesIn_Accounting / TotalMaleCount = 4/29

Prob_MaleCIS = CountOf_MalesIn_CIS / TotalMaleCount = 1/29

Prob_MaleEconomics_Finance = CountOf_MalesIn_Economics_Finance / TotalMaleCount = 4/29

Prob_MaleInternationalBusiness = CountOf_MalesIn_InternationalBusiness / TotalMaleCount = 2/29

Prob_MaleManagement = CountOf_MalesIn_Management / TotalMaleCount = 6/29

Prob_MaleOther = CountOf_MalesIn_Other/ TotalMaleCount = 4/29

Prob_MaleRetailing_Marketing = CountOf_MalesIn_Retailing_Marketing / TotalMaleCount = 5/29

Prob_MaleUndecided = CountOf_Males_Undecided/ TotalMaleCount = 3/29

2.3.2 Find the conditional probability of different majors among the female students of CMSU.

Based on the above shown table, Following are Conditional probabilities for males:

Prob_FemaleAccounting = CountOf_FemalesIn_Accounting / TotalFemaleCount = 3/33

Prob_FemaleCIS = CountOf_FemalesIn_CIS / TotalFemaleCount = 3/33

Prob_FemaleEconomics_Finance = CountOf_FemalesIn_Economics_Finance / TotalFemaleCount = 7/33

Prob_FemaleInternationalBusiness = CountOf_FemalesIn_InternationalBusiness / TotalFemaleCount = 4/33

Prob_FemaleManagement = CountOf_FemalesIn_Management / TotalFemaleCount = 4/33

Prob_FemaleOther = CountOf_FemalesIn_Other/ TotalFemaleCount = 3/33

Prob_FemaleRetailing_Marketing = CountOf_FemalesIn_Retailing_Marketing / TotalFemaleCount = 9/33

Prob_FemaleUndecided = CountOf_Females_Undecided/ TotalFemaleCount = 0/33

2.4. Assume that the sample is a representative of the population of CMSU. Based on the data, answer the following question:

2.4.1. Find the probability That a randomly chosen student is a male and intends to graduate.

P(A) : Probability That a randomly chosen student is a male
P(B) : Probability That a randomly chosen student intends to graduate
= 1 - (Probability That a randomly chosen student is undecided)
= 1 – 3/29 = 26/29
As per the addition rule, when the events are mutually exclusive,

P(AUB) = P(A)+P(B)

Applying the formula, we get :

P(AUB) = 29/62 * (26/29) = 26/62

2.4.2 Find the probability that a randomly selected student is a female and does NOT have a laptop.

Following is the Created contingency table of Gender and Computer :

Computer	Desktop	Laptop	Tablet	All
Gender				
Female	2	29	2	33
Male	3	26	0	29
All	5	55	2	62

P(A) : Probability That a randomly chosen student is a Female
P(B) : Probability That a female doesn't have a laptop
P(A∩B) : Probability that a randomly selected student is a female and does NOT have a laptop.

From the contingency table shown above, it's clear that there are 4 females out of 33 who don't have a laptop

Hence, **P(A∩B)** = 4/62

2.5. Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question:

2.5.1. Find the probability that a randomly chosen student is either a male or has full-time employment?

P(A) : Probability that a Randomly chosen Student Is Male = 29/62

P(B) : Probability that a Randomly chosen Student has full-time employment = 10/62

P(A∩B) : Probability that a randomly chosen student is a male AND has full-time employment = 7/62

P(AUB) : Probability that a randomly chosen student is a male OR has full-time employment

Applying the Addition Rule, we get :

As per the addition rule, when the events aren't mutually exclusive,

P(AUB) = P(A)+P(B)–P(A∩B)

Applying the formula, we get :

P(AUB) = 29/62 + 10/62 – 7/62 = 32/62

2.5.2. Find the conditional probability that given a female student is randomly chosen, she is majoring in international business or management.

P(A) : Probability that a Randomly chosen Student Is Female = 33/62

P(B) : Probability of Majoring In International Business Or Management

= Probability of Majoring In International Business + Probability of Majoring In Management

= 4/62 + 4/62 = 8/62

P(B/A) : Conditional Probability of Majoring In International Business Or Management Given the Randomly chosen student is Female

When events are not independent,

P(B/A) = P(A∩B) / P(A)

Applying the formula, we get :

P(B/A) = (8/62) / (33/62) = 8/33

2.6. Construct a contingency table of Gender and Intent to Graduate at 2 levels (Yes/No). The Undecided students are not considered now and the table is a 2x2 table. Do you think the graduate intention and being female are independent events?

Following is the Created contingency table of Gender and Intent to Graduate :

Grad Intention	No	Yes	All
Gender			
Female	9	11	20
Male	3	17	20
All	12	28	40

In order to prove that graduate intention and being female are independent events,

Let us assume that:

The graduate intention and being female are independent events.Then the Multiplication Rule for Independent events must also hold true.Hence.

Prob_RandomStudentIsFemaleANDProb_RandomStudentHavingGradIntention = Prob_RandomStudentIsFemale * Prob_RandomStudentHavingGradIntention

R.H.S = 20/40 * 28/40 = 7/20 = 0.35

Now, calculating **L.H.S using contingency table ,**

Prob_RandomStudentIsFemaleANDProb_RandomStudentHavingGradIntention = 11/40 = 0.275

Since L.H.S ≠ R.H.S , **We can conclude that graduate intention and being female are not independent events.**

2.7. Note that there are four numerical (continuous) variables in the data set, GPA, Salary, Spending, and Text Messages.

Answer the following questions based on the data

2.7.1. If a student is chosen randomly, what is the probability that his/her GPA is less than 3?

CountOfGpaLessThan3 has been calculated using python filter condition on the dataframe:

Length of Dataframe subset where GPA is less than 3 = 17

Now,

Prob_ChosenStudentHasGPAlessThan3 = CountOfGpaLessThan3/TotalCount = 17/62

2.7.2. a) Find the conditional probability that a randomly selected male earns 50 or more.

Entries where **male earns 50 or more** & where a **female earns 50 or more** have been calculated using python filter condition and the contingency table has been created:

Salary	50.0	52.0	54.0	55.0	60.0	65.0	70.0	78.0	80.0	All
Gender										
Female	5	0	0	5	5	0	1	1	1	18
Male	4	1	1	3	3	1	0	0	1	14
All	9	1	1	8	8	1	1	1	2	32

Using the output of contingency table, I have calculated marginal probabilities using another simplified table :

	Salary>=50	**Salary<50**	**All**
Female	18	15	33
Male	14	15	29
All	32	30	62

P(A) : Probability that the randomly selected person earns 50 or more

P(B) : Probability that a randomly selected person is male

P(A/B) : Probability that the selected person earns 50 or more Given the person is male.

P(A/B) = P(A∩B) / P(B) ,

P(A∩B) : Probability that the selected person earns 50 or more AND the person is male

From the above shown table, it is clear that there are total 14 males Earning 50 Or more , among a total of 62 people.

Hence, P(A∩B) = 14/62

P(B) = Number of males / Total Number of undergraduate students attending CMSU

Applying the formula, we get :

P(A/B)= (14/62) / (29/62) = 14/29

2.7.2 b) Find the conditional probability that a randomly selected female earns 50 or more.

Similarly, for

P(A/B) : Probability that the selected person earns 50 or more Given the person is female.

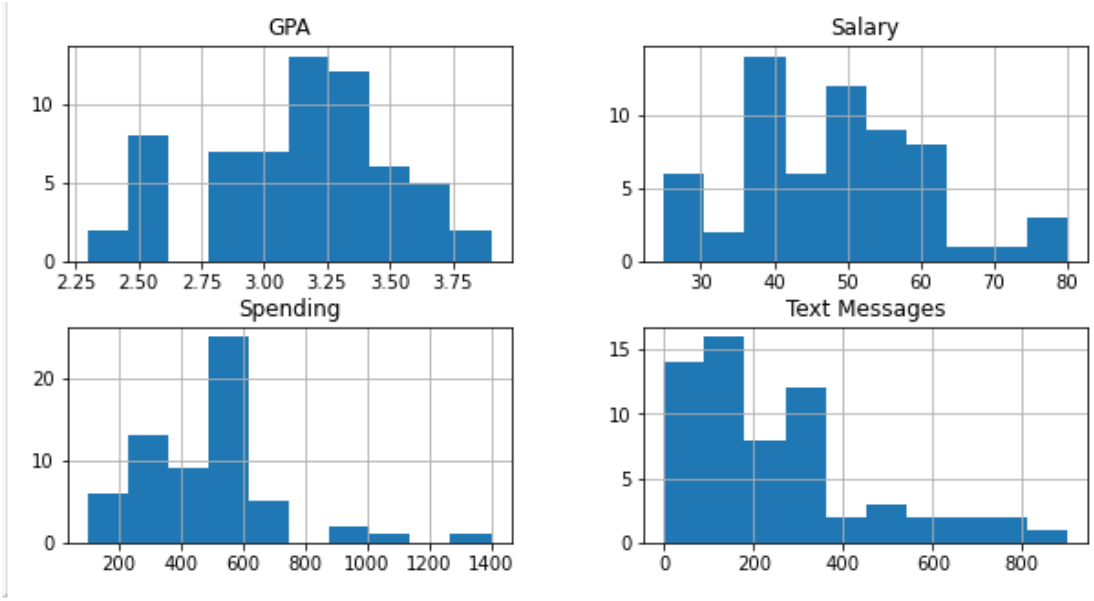
P(A/B) = (18/62) / (33/62) = 18/33

2.8.a) Note that there are four numerical (continuous) variables in the data set, GPA, Salary, Spending, and Text Messages. For each of them comment whether they follow a normal distribution.

The properties of a random sample having normal distribution are:

1. The data distribution has to be symmetric, It cannot be left or right skewed.
2. The mean, mode, median are all equal.
3. Half of the distribution count is less than the mean & the other half is greater than the mean.

Having said that, following are the histograms plotted for all four continuous variables:



OBSERVATIONS FROM HISTOGRAM:

1. '**GPA'** follows close to normal kind of distribution , appears a bit left-tailed/negatively skewed. The peak is around 3.10 , but the distribution extends backwards to lower values of left side, as compared to the higher values of right side.

2. '**Salary'** follows close to normal kind of distribution , appears a bit right-tailed/positively skewed. The peak is around 36 , but the distribution extends further to higher values of right side, as compared to the lower values of left side.

3. '**Spending'** follows a positively skewed distribution , is right-tailed/positively skewed. The peak is around 440 , but the distribution extends further to higher values of right side, as compared to the lower values of left side.

4. '**Text Messages**' follows a positively skewed distribution , is right-tailed/positively skewed. The peak is around 190 , but the distribution extends further to higher values of right side, as compared to the lower values of left side.

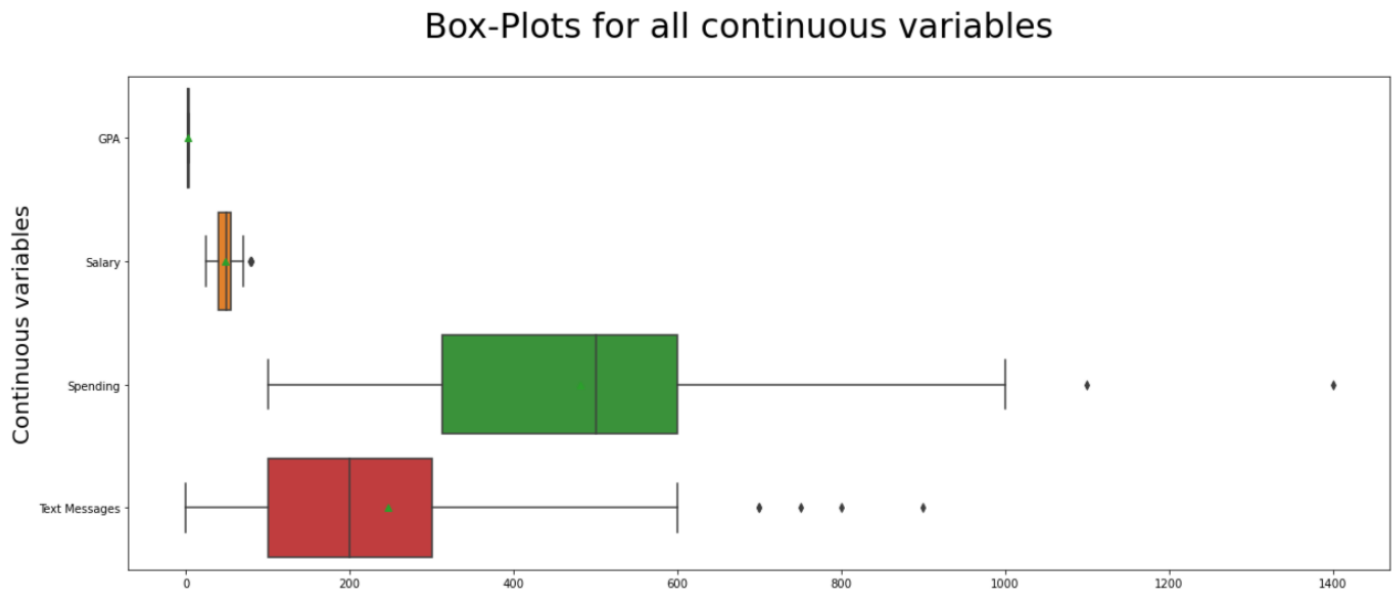
	Skewness
GPA	-0.314600
Salary	0.534701
Spending	1.585915
Text Messages	1.295808

Same can be concluded upon calculation of **Skewness** parameter values for all 4 columns.

2.8.b)Write a note summarizing your conclusions for this whole Problem 2.

1. Negative skewness in '**GPA'** shows that very few undergraduates are scoring GPA below average.This implies that most of the students are performing well.

2. There is huge difference between **'Salary'** and '**Spending'** of students.Also, the salary obtained by the students is almost same in numbers, but their spending pattern is quite different comparatively.



	GPA	Salary	Spending	Text Messages
count	62.000000	62.000000	62.000000	62.000000
mean	3.129032	48.548387	482.016129	246.209677
std	0.377388	12.080912	221.953805	214.465950
min	2.300000	25.000000	100.000000	0.000000
25%	2.900000	40.000000	312.500000	100.000000
50%	3.150000	50.000000	500.000000	200.000000
75%	3.400000	55.000000	600.000000	300.000000
max	3.900000	80.000000	1400.000000	900.000000

Problem #3

An important quality characteristic used by the manufacturers of ABC asphalt shingles is the amount of moisture the shingles contain when they are packaged. Customers may feel that they have purchased a product lacking in quality if they find moisture and wet shingles inside the packaging. In some cases, excessive moisture can cause the granules attached to the shingles for texture and colouring purposes to fall off the shingles resulting in appearance problems. To monitor the amount of moisture present, the company conducts moisture tests. A shingle is weighed and then dried. The shingle is then reweighed, and based on the amount of moisture taken out of the product, the pounds of moisture per 100 square feet is calculated. The company would like to show that the mean moisture content is less than 0.35 pound per 100 square feet.

The file (A & B shingles.csv) includes 36 measurements (in pounds per 100 square feet) for A shingles and 31 for B shingles.

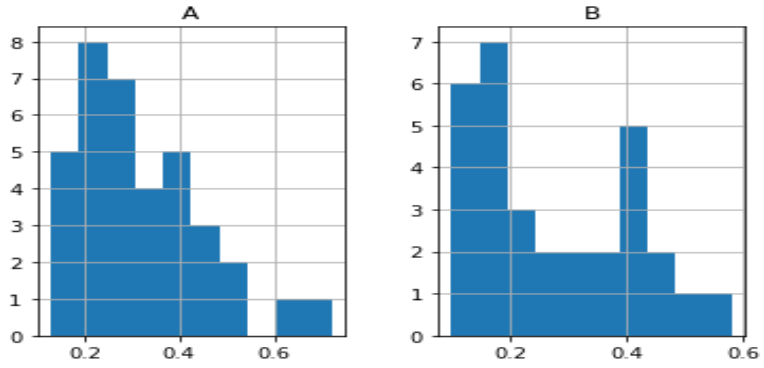
3.1 Do you think there is evidence that means moisture contents in both types of shingles are within the permissible limits? State your conclusions clearly showing all steps.

Analyzing the dataset using **describe**() and by plotting Histogram & box-plot, we get the following :

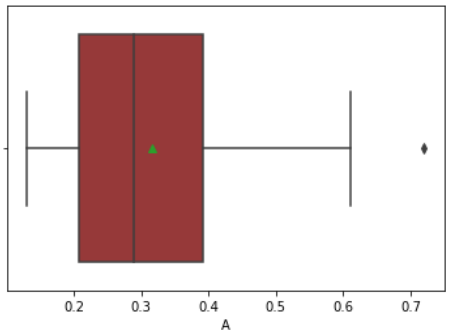
```
data.describe()
```

	A	B
count	36.000000	31.000000
mean	0.316667	0.273548
std	0.135731	0.137296
min	0.130000	0.100000
25%	0.207500	0.160000
50%	0.290000	0.230000
75%	0.392500	0.400000
max	0.720000	0.580000

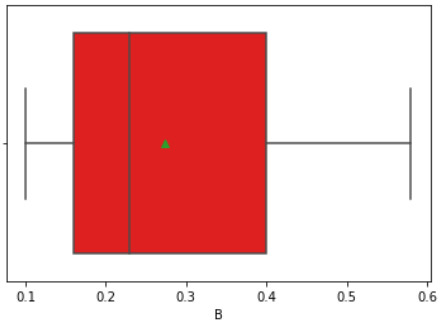
```
data.hist()
plt.show()
```




```
sns.boxplot(x=data['A'],color='brown',showmeans =True);
```




```
sns.boxplot(x=data['B'],color='red',showmeans =True);
```



- Mean and Median Values of each sample are not much different.
- The 'A' sample data looks more symmetrically distributed, whereas the 'B' sample data looks right skewed.
- But since the population standard deviation (Sigma) is unknown, we have to use a Tstat test.
- Since we need to check if mean moisture contents in both types of shingles are within the permissible limits individually, we'll perform individual **1 sample T-Tests** using **ttest_1samp** method from scipy library for both "A" & "B".

FOR 'A':

Step 1: Define null and alternative hypotheses

- Null hypothesis states that mean moisture content μ =< 0.35, Ho => μ =< 0.35
- Alternative hypothesis states that the mean moisture content μ > 0.35, Ha => μ > 0.35

Step 2: Decide the significance level

- The level of significance (Alpha) = 0.05.

Step 3: Identify the test statistic

- tstat: -1.4735046253382782
- p-value for one-tail = p-value/2 : 0.07477633144907513

Hence, We have no evidence to reject the null hypothesis since p value > Level of significance

FOR 'B':

Step 1: Define null and alternative hypotheses

- Null hypothesis states that mean moisture content μ =< 0.35, Ho => μ =< 0.35
- Alternative hypothesis states that the mean moisture content μ > 0.35, Ha => μ > 0.35

Step 2: Decide the significance level

- The level of significance (Alpha) = 0.05.

Step 3: Identify the test statistic

- tstat : -3.1003313069986995
- p-value for one-tail = p-value/2 : 0.0020904774003191826

Hence, We have evidence to reject the null hypothesis since p value < Level of significance

3.2 Do you think that the population mean for shingles A and B are equal? Form the hypothesis and conduct the test of the hypothesis. What assumption do you need to check before the test for equality of means is performed?

- We use the **scipy.stats.ttest_ind** method from scipy library to calculate the t-test for the means of TWO INDEPENDENT samples of scores given the two sample observations. This function returns t statistic and two-tailed p value.

Step 1: Define null and alternative hypotheses

- Null hypothesis states that the population mean moisture content for A & B are equal
- Alternative hypothesis states that the mean moisture content for A & B are not equal
- $H0$: μA - μB = 0 i.e., $\mu A = \mu B$
- HA: μA - $\mu B \neq 0$ i.e., $\mu A \neq \mu B$

Step 2: Decide the significance level

- The level of significance (Alpha_level) = 0.05 , as this is the default value of Confidence-level when not provided. The population standard deviation is also not known.

Step 3: Identify the test statistic

Tstat: 1.2896282719661123

p-value: 0.2017496571835306

Since p-value> Alpha_level,

We do not have enough evidence to reject the null hypothesis in favour of alternative hypothesis.
Hence, We conclude that the population mean for shingles A and B are equal.

ASSUMPTIONS MADE:

- This is a two-sided test for the null hypothesis that 2 independent samples have identical average (expected) values. This test assumes that the populations have identical variances.
- Before the test for equality of means is performed, we are going to assume that the variance is equal and then compute the necessary statistical values.
- The samples are assumed to be following a normal distribution.